Statistical Analysis of Clinical Phenotype and Patient Survival in Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies

Dean Marchiori*

Abstract

A recently described, rare genetic condition known as Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies (NEDMABA) has been identified in children with bi-allelic loss-of-function variants in *SMPD4*. The progression of this condition is not well understood with the limited case reports described so far exhibiting a severe and clinically diverse phenotype. A gap exists in the understanding of associations present in the heterogenous features of the clinical phenotype, and the expected survival probabilities of affected individuals. This is driven in part to the paucity of analysis-ready data on reported cases. This research aims to collate and standardise available case reports, to analyse and identify meaningful clusters in the clinical phenotype, and to quantify the survival probability for children with NEDMABA. To overcome the challenge of multidimensional data on very few subjects, we employ Multiple Correspondence Analysis (MCA) as a dimension reduction technique, which is then subject to cluster analysis and interpretation. To quantify more accurate survival probabilities, Kaplan-Meier estimation is formulated to account for censoring in the data. The analysis correctly identified the classic phenotype for this condition, as well as one other distinct feature cluster relating to findings of vocal cord paralysis and respiratory failure. Several additional outlying and isolated clinical features were also identified. The survival probability for those affected was found to decline sharply in early infancy with median survival of 150 days, but with some surviving as long as 12.5 years. This wide range of outcomes is provisionally associated with variant type however interpretation of these results are guarded based on very low sample sizes. This analysis represents the first of its kind to help describe associations and trajectories of individuals with this condition, despite challenges with sparse and inconsistent data. This analysis can provide clinicians and genetic counsellors with better information to aide in decision making and support for families with this rare condition.

Keywords: NEDMABA, SMPD4

1. Introduction

Neurodevelopment disorders are a diverse and heterogeneous group of conditions impacting the development of the nervous system, brain function, physical development and learning ability. A recently identified condition known as Neurodevelopmental Disorder with Microcephaly, Arthrogryposis, and Structural Brain Anomalies (NEDMABA) (MIM:618622) has been described in children with bi-allelic loss-of-function variants in *SMPD4* (Magini et al., 2019). Sphingomyelinases such as *SMPD4* play an important cellular role by hydrolyzing sphingomyelin into ceramide and phosphorylcholine. *SMPD4* specifically encodes one of 4 neutral sphinogmyelinases, nSMase3 (MIM: 610457). So far, less than 50 cases of this rare disorder have been reported in the literature. Clinical phenotype data of these cases is largely heterogeneous with severe

---

*Corresponding author

   Email address: `deanmarchiori@gmail.com` (Dean Marchiori)

neurological complications and early demise a key feature. This makes the identification, diagnosis and ongoing management of cases a challenge for patients, their families and medical practitioners.

Ravenscroft et al. (2021) performed a study over 190 probands and identified a novel missense variation in SMPD4 which at the time of whole exome sequencing was not well described in the literature. The impacted family from Melbourne exhibited features involving arthrogryposis multiplex congenita, complex brain malformations, small for gestation age and hypoplasia of the corpus callosum. In two of the three related cases were additional features of microcephaly, congenital encephalopathy, cerebellar malformation and hypoplasia and hypomyelination.

A detailed study by Magini et al. (2019) involved 12 unrelated families with 32 individuals (21 with detailed clinical information). Presentations were of microcephaly, simplified gyral pattern of the cortex, hypomyelination, cerebellar hypoplasia, congenital arthrogryposis, and early fetal/postnatal demise. Despite this being the largest cohort studied, the clinical features and survival times among participants varied greatly. Three missense changes were noted in the study with affected children in these families often showing a milder presentation suggestive of possible residual function. In these cases individuals were able to develop independent motor skills, have mild intellectual disability and arthrogryposis without evidence of simplified gyral patterns on brain MRI. Other patients with truncating variants are shown to have more severe presentations, while a range of additional significant clinical features were reported involving dysmorphic facial features, seizure, vocal cord paralysis and hearing impairment.

A case study from China is described in Ji et al. (2022) involving a girl presenting in infancy with intrauterine growth restriction, microcephaly, postnatal developmental delay, arthrogryposis, hypertonicity, seizure, and hypomyelination on brain magnetic resonance imaging. The authors report parallels with two cases showing the same homozygous null variant. An individual reported in Monies et al. (2019) presented with distinct symptoms of brain atrophy and skeletal dysplasia whereas the case in Magini et al. (2019) with the same variant exhibited more typical clinical features.

A recent study by Bijarnia-Mahay et al. (2022) presents the case of a 22-month old girl presenting with the typical phenotype of neurodevelopmental delay, prenatal onset growth failure, arthrogryposis, microcephaly and brain anomalies including severe hypomyelination, simplified gyral pattern and hypoplasia of corpus callosum and brainstem. Notably, there was also additional non-typical clinical findings of nystagmus and visual impairment secondary to macular dystrophy and retinal pigment epithelial stippling at posterior pole.

Further work to collate and analyse data relating to NEDMABA is challenging. While Magini et al. (2019) cataloged a detailed clinical phenotype data set as a supplementary data set to their study, these data are not suitable for statistical analysis as presented. Many of the clinical features are represented as free text descriptions and a variety of non-standard terminologies are applied between cases, making machine interpretation of the data difficult. Other studies (Ji et al., 2022; Bijarnia-Mahay et al., 2022) present only written case reports with some tabulated summaries. While larger studies (Ravenscroft et al., 2021; Monies et al., 2019) focus more on genetic data, with less detail included on the clinical presentation of individuals in the cohort.

To overcome these data challenges, Wickham (2014) proposes a data structure know as 'tidy' data which organises each observation in a row, with each feature as its own column and each value in just one cell. In this case detailed text-based data can be transformed to high dimensional binary indicators of key clinical features. This structure is conducive to effective statistical analysis and integrates deliberately with the *tidyverse* (Wickham et al., 2019) collection of packages for data analysis in R (R Core Team, 2022).

Further analysis to better describe rare genetic conditions was explored in Díaz-Santiago et al. (2020) in the context of large scale genotype-phenotype analysis. The authors argue patients with rare disorders (often in small samples) often present with varied symptoms that do not match exactly with the described phenotype. When considering this situation of low sample size data with many features, methods in the field of multivariate statistical analysis are commonly deployed. Here the aim is to find substructure in the data, or a simple representation of the multi-dimensional space. Methods, such as Multiple Correspondence

Analysis (Le Roux and Rouanet, 2010) are appropriate for high dimensional data comprised of binary indicators. This technique is cited in many studies in the application of dimension reduction and clustering of comorbidities and phenotypes from disease (Han et al., 2018; Costa et al., 2013).

In terms of patient survival, methods in statistical survival analysis are well suited and commonly used in the field of genomics (Chen et al., 2014). These methods in particular provide a mechanism for dealing with right-censoring, a common feature of clinical studies where the clinical outcome of interest may not be known by the end of the study period.

An open research question exists around the clinical pathway and survival time of children exhibiting bi-allelic loss-of-function variants in *SMPD4*. With current research highlighting a diverse and heterogeneous phenotype, the relationship and correlations between these diverse features is not well understood. Furthermore, the survival time of affected children has not been extensively analysed despite early-demise featuring as a severe outcome in many cases. While a connection has been highlighted between some missense variants and a milder presentation with longer survival (Magini et al., 2019), this hypothesis has not been analysed further.

This research has three key aims. First, to collate and transform the variety of early case reports and studies on clinical phenotype data for this novel variant into an analysis-ready *tidy* data set. Secondly, to conduct analysis on the associations between commonly reported clinical features. Finally, to statistically quantify the expected survival time of children with NEDMABA based on the current case reports.

This will be the first in-depth analysis of this novel variant, which aims to produce statistical findings to better understand this newly described condition. This research will assist clinicians understand the various presentations of this challenging new condition. In addition, genetic counselling of families with affected children will benefit from enhanced analysis on outcomes of existing reported cases.

2. Methods

2.1. SMPD4 Data Package

The data contained in Magini et al. (2019) "Summary of SMPD4-Related Clinical Phenotype" is an excel spreadsheet tabulating each of the 21 individuals from the study with clinical details recorded. The file has 21 columns (one for each individual), and 64 rows (one for each phenotype or clinical remark).

The data were read into R (R Core Team, 2022) without any changes, so as to preserve the reproducibility of the data transformation steps. The data were transposed to ensure it could be presented as *tidy* formatted data (Wickham, 2014). This requires each variable (clinical phenotype) forms a column; each observation (individual) forms a row and each type of observational unit forms a table (every cell has just one item). This resulted in a long-formatted data set of 21 rows (individuals) and 64 columns (clinical features). In many cases, key clinical information was entered as free-text descriptions, which rendered any attempt of meaningful analysis impractical (Table 1). In these cases, the text was tokenised by separating the list of clinical observations at each comma and forming a binary indicator column noting its presence '1' or absence '0' (Table 2).

Table 1: Example of non-tidy free-text descriptions of features

	Family 1- Individual 1	Family 1- Individual 4
Facial dysmorphisms	short palpebral fissures, large ears, simple helices, smooth philtrum, thin lips, bilateral simian creases	short palpebral fissures, receding forehead, thin upper lip

Table 2: Example of tidy formatted data where individuals are transposed into rows and text into binary indicators

id	short palpebral fissures	large ears	simple helices	receding forehead	...
Family 1-Individual 1	1	1	1	0	...
Family 1-Individual 4	1	0	0	1	...

In the case where two variables were formed from features that are considered to be synonymous, these were merged into one indicator column to prevent duplication e.g. {bilateral_cleft_lips, bilateral_cleft_lip, cleft_lip_b_l} -> {bilateral_cleft_lip}.

Data type conversion and categorical level standardisation was performed to ensure the data were in consistent and appropriate data types. For example, 'Gender' was not consistently coded, and 'Birth Weight' was encoded as a text string rather than a more useful numeric format. (Table 3).

Table 3: Example of inconsistent coding or sub-optimal data types

Gender	Birth Weight
male	2175 grams (- 2.5 SD)
female	2045 g (-3 SD)
Female	2300 gram (-2 SD)
Female fetus	n.a.

Table 4: Example of consistent and appropriate coding or data types

Gender	Birth Weight (g)
male	2175
female	2045
female	2300
female	NA

This format was preserved and other case studies identified in the literature (Ravenscroft et al., 2021; Monies et al., 2019; Bijarnia-Mahay et al., 2022; Ji et al., 2022) were manually entered to conform to this template to allow the data to be combined for further analysis.

The final data set consisted of 28 observations (one per individual) and 152 variables (one per clinical feature). These variables were comprised of qualitative text descriptions, binary indicators of features as described above and other fields including background information on the patient and various bio markers. A full list of this data set is available in Table 10. These data sets were compiled into a publicly available R Package called SMPD4 (smp) in order to allow for reproducibility and sharing.

2.2. Multiple Correspondence Analysis

The data on subjects introduced above is subset to include only variables that can be transformed into binary indicator variables that represent the presence or absence of a given clinical feature. Only those features that appeared in more than one case were included to minimise the influence of isolated features in this analysis.

Dimension reduction on this wide data set was performed using Multiple Correspondence Analysis (MCA) (an analogy to Principle Component Analysis (PCA) for categorical data) (Le Roux and Rouanet, 2010) in order to arrive at a set of meaningfully small dimensions that account for most of the variation in the data.

We let these data be $\mathbf{X}$ which is comprised of a set of individuals I and a set of features Q such that the q_{th} feature has K_q levels. The sum of all categories $K = \sum_{q=1}^{Q} K_q$ defines the dimensionality of $\mathbf{X}$ as an $I \times K$ matrix. This resulted in a data set $X_{28 \times 61}$ indicator matrix of 61 clinical features across 28 individuals. Taking $\delta_{ik} = 1$ if subject i has feature k and $\delta_{ik} = 0$ if the subject does not, we are left with the completely disjunctive table $\mathbf{X} = I \times K$ of $\{0, 1\}$. Letting the sum of all entries of $\mathbf{X}$ be N, we have $\mathbf{Z} = N^{-1}\mathbf{X}$. We can introduce two diagonal matrices $\mathbf{D_r} = diag(\mathbf{r})$ and $\mathbf{D_c} = diag(\mathbf{c})$ where $\mathbf{r}$ and $\mathbf{c}$ are the vectors of row sums and column sums of $\mathbf{Z}$ respectively.

Computing MCA involves taking the Singular Value Decomposition:

$$\mathbf{M} = \mathbf{D_r}^{-\frac{1}{2}}(\mathbf{Z} - \mathbf{rc^T})\mathbf{D_c}^{-\frac{1}{2}} = \mathbf{P}\Delta\mathbf{Q^T}$$

where Δ are the singular values and $\Lambda = \Delta^2$ is the matrix of eigenvalues.

This results in the row and column factor scores respectively as:

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{P}\Delta$$

$$\mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{Q}\Delta$$

This computation of the Multiple Correspondence Analysis was computed using the *FactoMineR* R package (Lê et al., 2008).

The *contribution* of each category was computed for each of the principal axis per Le Roux and Rouanet (2010). This metric identifies the proportion of variance of that axis due to the point and is defined as:

$$Ctr_k = \frac{py_k^2}{\lambda}$$

Where p represents the point weighting, y is the coordinate relative to the principal axes of variance λ.

The coordinates on the first few principal axes were inspected for any outliers or unusual data.

2.3. Cluster Analysis

The MCA coordinates for the first 4 principal dimensions were subject to clustering in order to find groups of features that were internally homogeneous and externally heterogeneous. This was computed in R using the clValid package (Brock et al., 2008).

A range of clustering algorithms were compared including AGNES hierarchical agglomerative (Kaufman and Rousseeuw, 2009), k-means (Hartigan and Wong, 1979) and self-organising maps (SOM) (Kohonen, 2012). A sensible range of clusters from 2-5 were trialed for each clustering algorithm. The results for each method were evaluated using three measures of stability: Average proportion of non-overlap (APN), Average Distance (AD) and Average Distance Between Means (ADM) (Datta and Datta, 2003). Stability measures compare the full clustering result with a result based on dropping each of the columns one at a time. These metrics were selected to represent each data point (clinical feature) with a number of MCA dimensions. A good cluster solution is one that is robust and meaningful across all of these dimensions Figure 5. The optimal method and cluster number was selected based on a trade off between the best performing method using the evaluation metrics and a sensible partitioning of the data for this analysis Table 5.

2.4. Survival Analysis

To model survival probability for all subjects in the combined data, the data were subsetted to include *survival_time* which is either the number of days the individual survived for, or the age in days at last follow up, and *deceased* a numeric event indicator which is equal to 1 if the subject is deceased and 0 if the subject was still alive at the time the respective study had concluded. The Kaplan-Meier estimator (Kaplan and Meier, 1958) was used to estimate the survival function of the data.

The estimator is calculated as:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

where t_i is some event time, d_i represents the number of events (here deceased subjects) and n_i indicating the individuals known to have survived up to t_i

The baseline estimator of all individuals was calculated using the *survival* package in `R` (Terry M. Therneau and Patricia M. Grambsch, 2000). Next the Kaplan-Meier estimator was stratified by the type of genetic variation reported in the literature. A log-rank test was performed to detect differences in the survival curves using methods from Harrington and Fleming (1982), again implemented in the *survival* `R` package.

3. Results

3.1. MCA

The initial results of the MCA were inspected with a bi-plot and identified two subjects (17, 18) from the data that were clear outliers and sat significantly beyond the 95% confidence ellipse. These individuals represented two twins from Magini et al. (2019) who were described with a much milder phenotype and were subject to a missense variation in *SMPD4*. Further analysis of this influence is conducted below. For now, these subjects were removed to allow for more meaningful analysis.

Repeating MCA for the remaining subjects shows the first dimension ($\lambda_1 = 0.18$) accounts for 17.88% of the variance in the data with the first 4 dimensions of the MCA analysis accounting for 54% of the variance (Table 5).

Table 5: Eigenvalues and percentage of variance explained by the MCA principal axes

dimension	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.18	17.88%	17.88%
dim 2	0.14	13.81%	31.70%
dim 3	0.12	12.32%	44.01%
dim 4	0.10	9.66%	53.67%
dim 5	0.07	7.49%	61.16%
dim 6	0.06	5.97%	67.13%
dim 7	0.05	4.90%	72.03%
dim 8	0.04	4.20%	76.23%
dim 9	0.04	4.08%	80.31%
dim 10	0.03	3.47%	83.78%
dim 11	0.03	2.92%	86.70%
dim 12	0.02	2.49%	89.20%
dim 13	0.02	2.15%	91.34%
dim 14	0.02	1.82%	93.16%
dim 15	0.02	1.65%	94.81%
dim 16	0.01	1.12%	95.93%

dim 17	0.01	1.00%	96.93%
dim 18	0.01	0.93%	97.85%
dim 19	0.01	0.67%	98.53%
dim 20	0.01	0.63%	99.16%
dim 21	0.00	0.41%	99.57%
dim 22	0.00	0.31%	99.88%
dim 23	0.00	0.12%	100.00%
dim 24	0.00	0.00%	100.00%
dim 25	0.00	0.00%	100.00%

The *contribution* of the categories were ranked within each of the first four MCA dimensions (Figure 2). The first dimension is dominated by highly specific facial dysmorphisms such as epicanthal folds, short philtrum and low anterior hairline.

The second MCA dimension is categorised by feeding and respiratory dysfunction with vocal cord palsy, tracheostomy, tube feeding and GERD.

Examining the bi-plot of the first two principal axes Figure 3 illustrates the contrasting sources of variation between conditions relating to airway and swallowing function and predominantly facial features and structural brain anomalies such as as cerebellar hypoplasia, development delay.

The third dimension is characterised by atypical, complex features such as patent ductus arteriosus, respiratory infections, small for gestational age and further facial features. These are contrasted by the fourth principal dimension which is strongly categorised by respiratory distress including stridor and hypercapnia in the absence of typical presentations of simplified gyral pattern, intrauterine growth retardation and microcephaly.

3.2. Cluster Analysis

Cluster validation measures resulted in either 2 or 5 clusters as the optimal choice.

In order to partition the data into meaningful groupings, k-means with 5 clusters was selected. This reflects the optimal choice for both AD and FOM cluster metrics.

A scatter plot of the cluster results Figure 6 projected on just the first two dimensions highlights a core cluster of classic features of NEDMABA such as microcephaly, hypomyelination, arthrogryposis and feet deformity and structural brain anomalies. A separate cluster is formed of airway and feeding related conditions involving vocal cord paralysis, tracheostomy, tube feeding and enlarged ventricles.

Three other smaller satellite clusters exists where highly distinctive features were associated with a single individual or family. A full articulation of feature to cluster assignment is given in Table 9.

Table 6: Optimal clusters

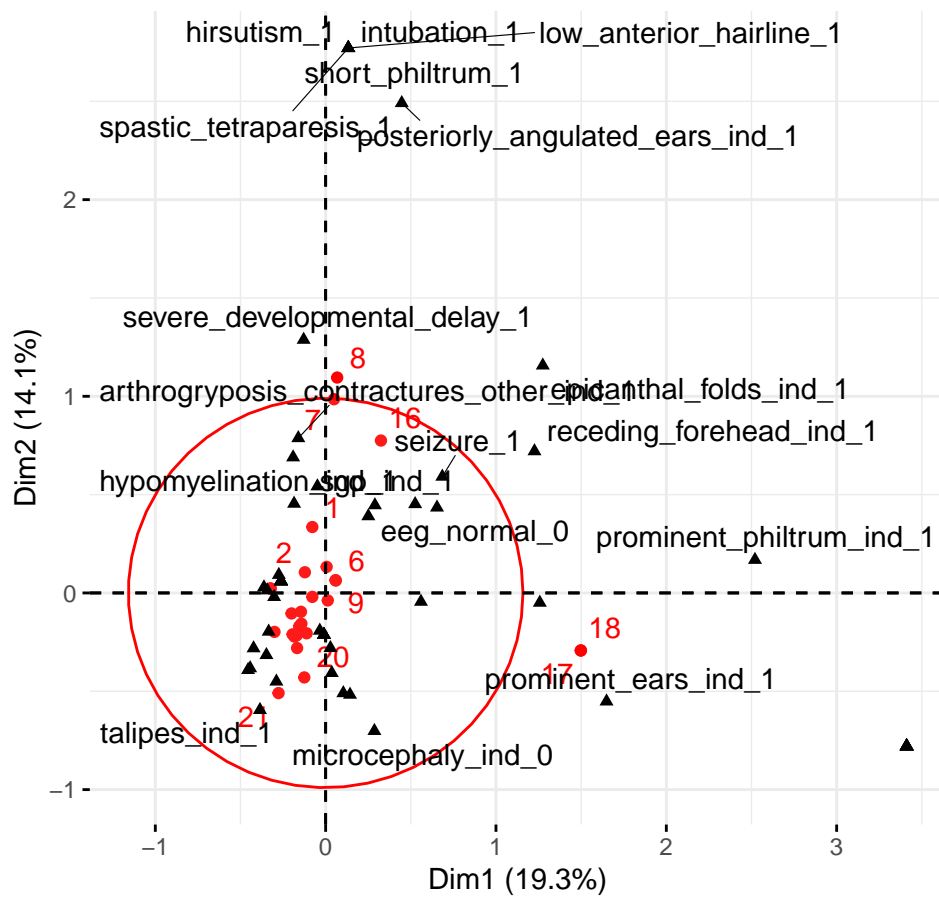| Metric	Score	Method	Clusters
APN	0.07	agnes	2
AD	1.38	kmeans	5
ADM	0.33	kmeans	2
FOM	0.65	kmeans	5

Figure 1: A bi-plot of the first two principal axes with grey circles indicating the individual's projected position and black triangles indicating the variables. A 95% confidence ellipse is plotted in red around the origin.
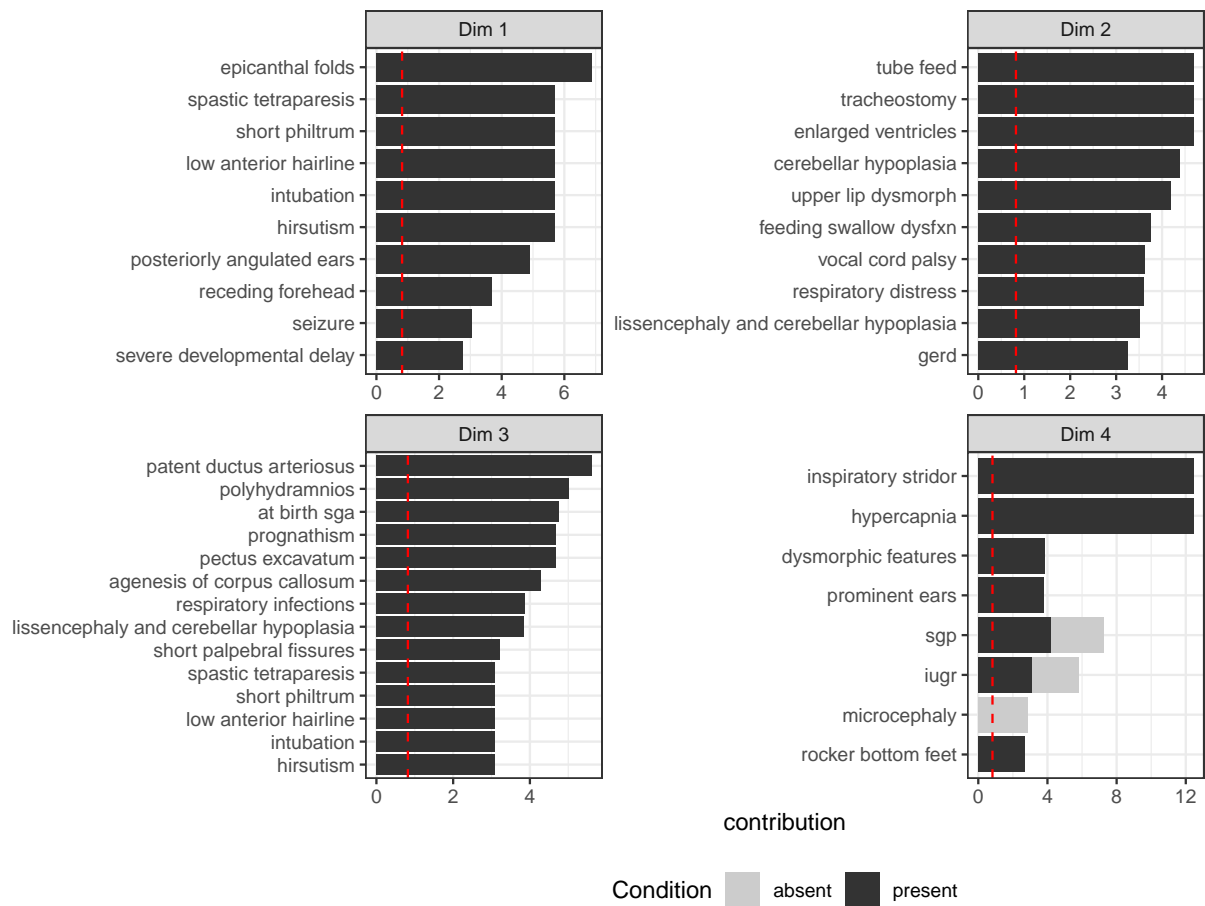
Figure 2: Top 10 feature categories from MCA analysis for the first four MCA dimensions. The red dashed line indicates the mean contribution value. Each feature may be present or absent as indicated by 1 or 0 respectively
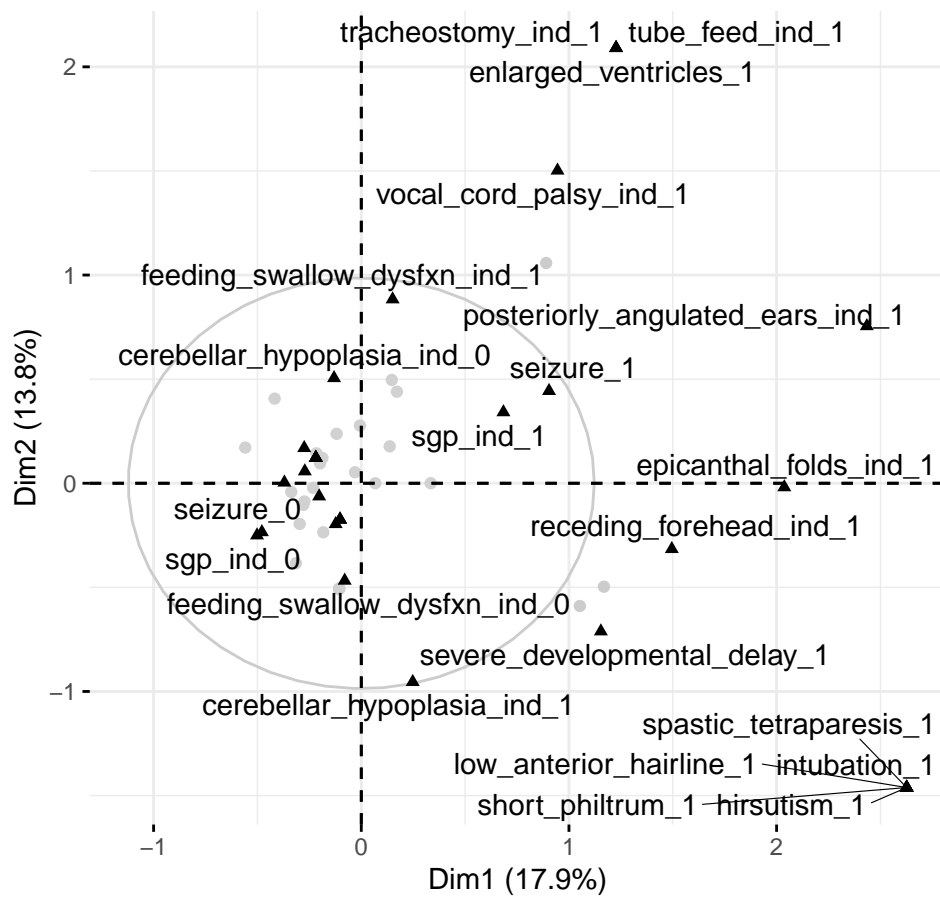
Figure 3: A bi-plot of the first two principal axes with grey circles indicating the individual's projected position and black triangles indicating the variables. A 95% confidence ellipse is plotted in grey around the origin.
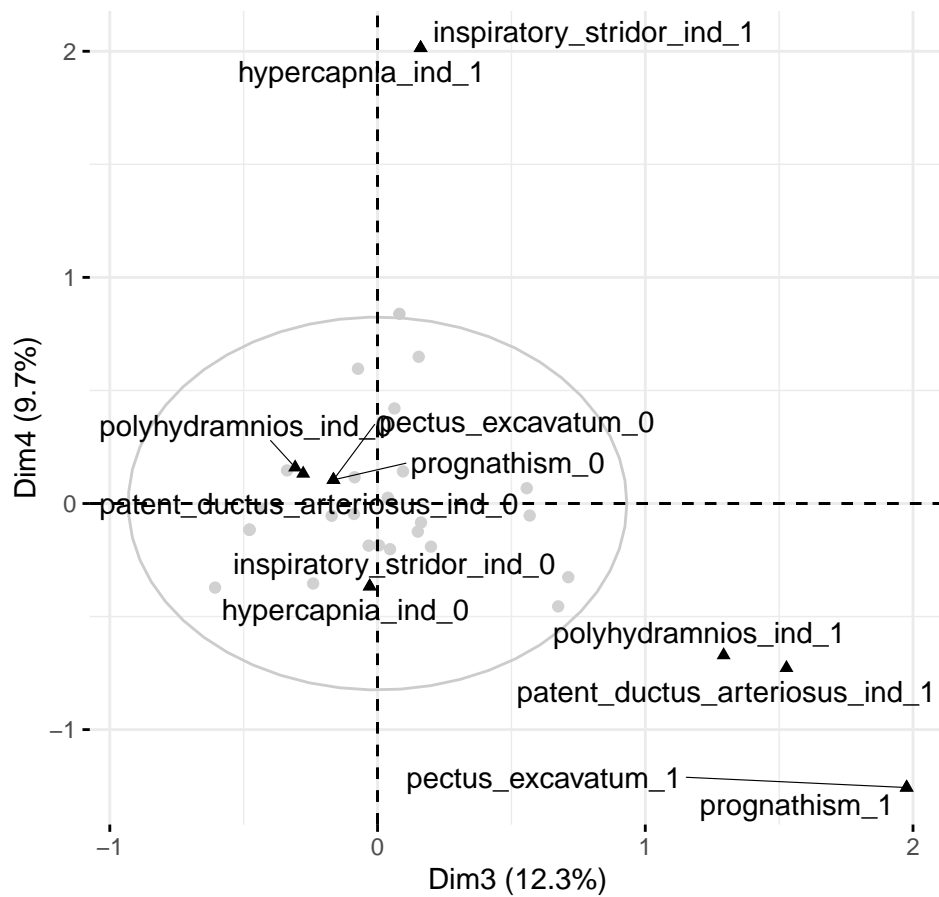
Figure 4: A bi-plot of principal axes 3 and 4, with grey circles indicating the individual's projected position and black triangles indicating the variables. A 95% confidence ellipse is plotted in grey around the origin.
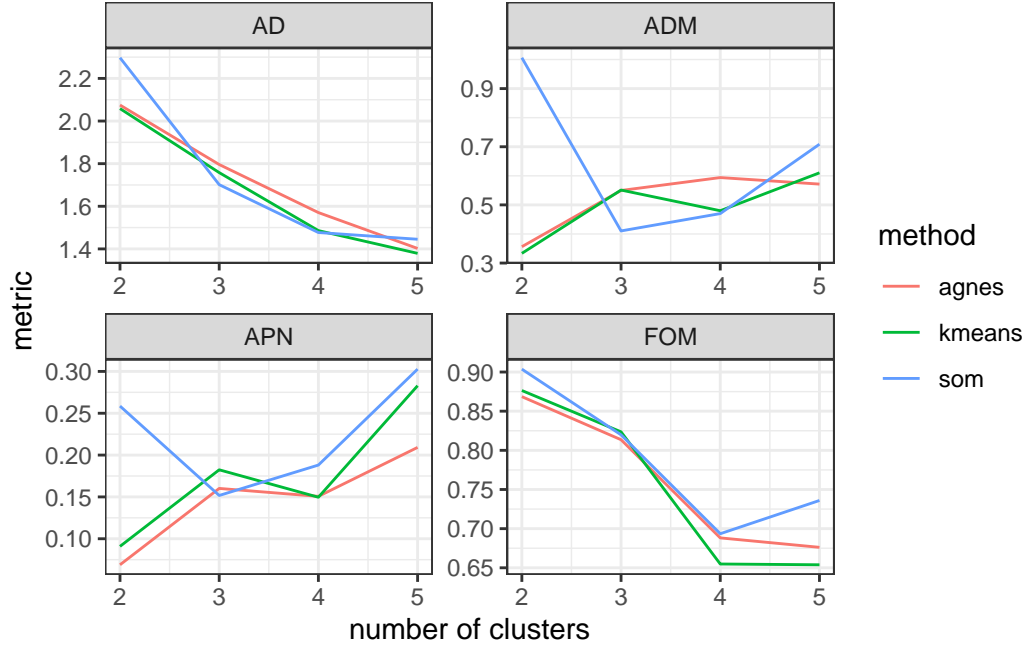
Figure 5: cluster validation analysis showing various metrics for cluster evaluation from 2 to 5 clusters. The most optimal number of clusters are selected by picking the lowest value from one or many of these comparable metrics.

3.3. Survival Analysis

The Kaplan-Meier estimator shows a range of survival times (0, 4562 days) with median survival time calculated as 150 days. At time 0 the survival probability was 0.82 (95% CI: 0.69, 0.98). Survival probability after one year at 0.403 (95% CI: 0.252, 0.644) and survival probability at close to ten years of 0.24 (95% CI: 0.11, 0.52).

Table 7: Survival probabilities derived from Kaplan-Meier estimation for the range of recorded survival times, including 95% Confidence Interval.

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
0.0	28	5	0	0.82	0.09	0.98	0.69
15.0	23	1	0	0.79	0.10	0.95	0.65
30.0	22	2	0	0.71	0.12	0.90	0.57
47.0	20	1	0	0.68	0.13	0.88	0.53
60.0	19	1	1	0.64	0.14	0.85	0.49
67.0	17	1	0	0.61	0.15	0.82	0.45
90.0	16	0	1	0.61	0.15	0.82	0.45
120.0	15	2	0	0.52	0.18	0.75	0.37
150.0	13	1	0	0.48	0.20	0.72	0.33
180.0	12	1	0	0.44	0.22	0.68	0.29
240.0	11	1	0	0.40	0.24	0.64	0.25
270.0	10	0	1	0.40	0.24	0.64	0.25
365.0	9	0	1	0.40	0.24	0.64	0.25
750.0	8	1	0	0.35	0.27	0.60	0.21
1095.0	7	1	0	0.30	0.31	0.56	0.16
1825.0	6	0	1	0.30	0.31	0.56	0.16

12

time	n.risk	n.event	n.censor	surv	std.err	upper	lower
2190.0	5	1	0	0.24	0.39	0.52	0.11
3285.0	4	0	2	0.24	0.39	0.52	0.11
4380.0	2	1	0	0.12	0.81	0.59	0.02
4562.5	1	1	0	0.00	Inf	NA	NA

A comparative box plot of survival times by variant type shows significant outliers for missense+splice and missense variations. However, even within missense homozygous variants the survival time varies greatly. Other missense variants, such as heterozygous missense with gene deletion do not show the same increased survival time as noted by Magini et al. (2019).
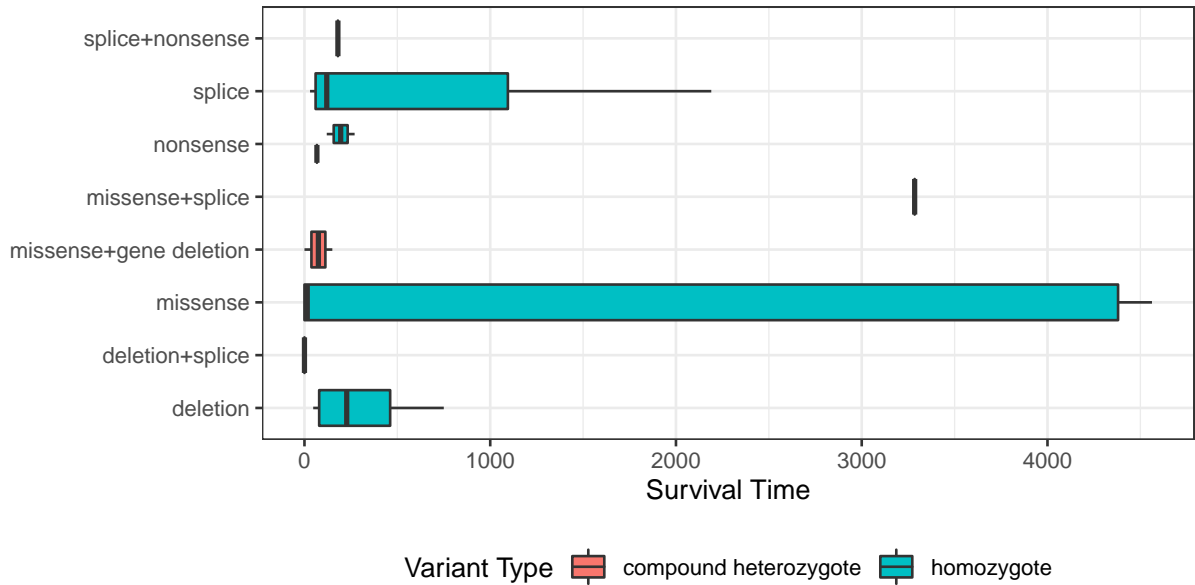
A log-rank test was carried out to compare the survival curves of the various gene variants and showed a significant difference in the stratified survival curves ($\chi^2 = 14.8, p = 0.04$) however due to the small sample size, several variants had expected counts less than 5 which would create a violation in assumptions for the test.

4. Discussion

The use of Multiple Correspondence Analysis confirmed the heterogeneous variation in the data with the first four principal axes being required to explain more than 50% of the variation in the data set. The examination of key contributing factors, as well as a visual inspection of the data projected into this lower dimensional space indicate the primary sources of variation in the clinical phenotype are explained by a few individuals with very different clinical features. This has confounded the typical or classical presentation.

The cluster analysis does not aim to group individuals that are alike, but rather the clinical features themselves and how these contrasting clusters account for variation in the data. This highlights common sources of variation in the clinical phenotype of the condition itself, rather than typical coincident features of any one child.

After excluding two outlying, remarkable cases this analysis clearly identifies a core cluster of the hallmark features of microcephaly, hypomyelination, arthrogryposis, simplified gyral pattern and developmental delay.
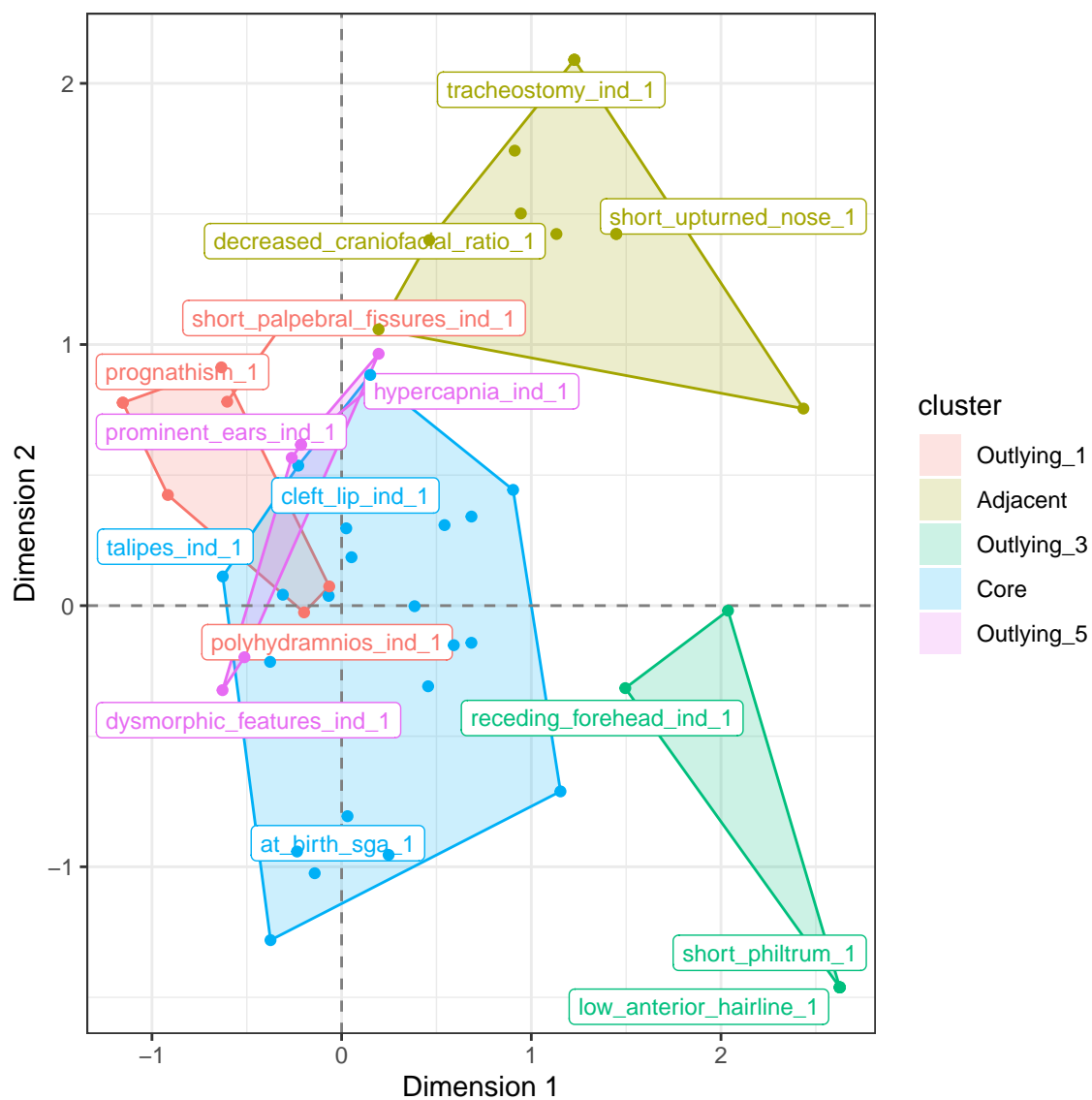
13

Figure 6: A projection of features on the first two principal components with k-means cluster assignment indicated by colour. A random sample of three labels per cluster are printed to provide a representation of each cluster's characteristics.
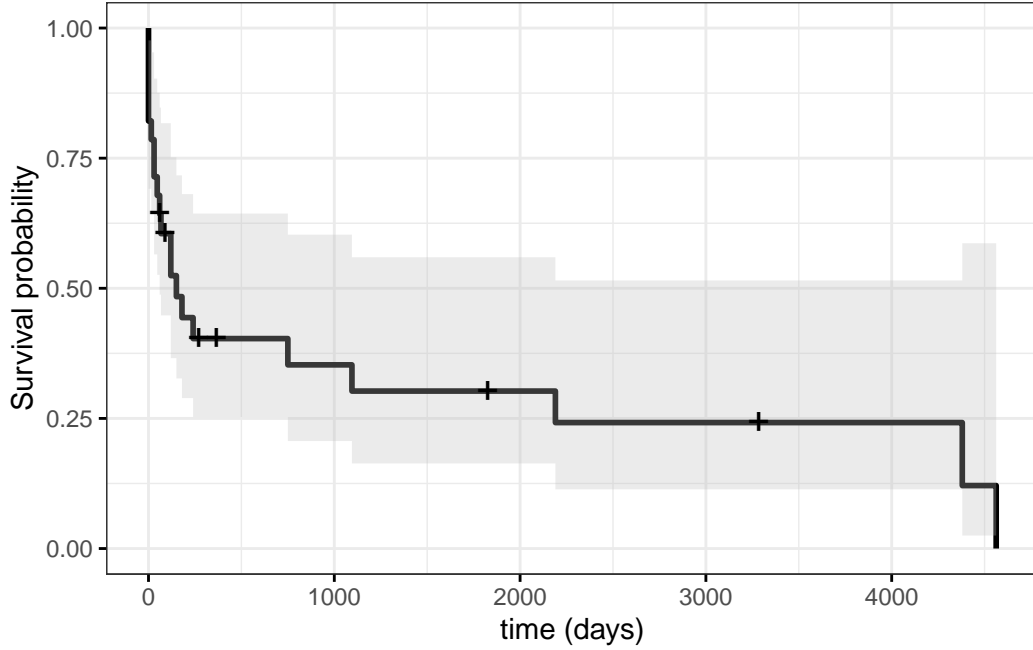
Figure 7: Kaplan Meier curve with the black line representing the survival probability and with grey shading indicating the 95% confidence interval. Cross icons represent right-censoring

Of significant note is the presence of an adjacent cluster of related features that are not as well described in The Human Phenotype Ontology (Köhler et al., 2021) as shown in Table 8. We identify a prolific feature of 'respiratory distress' as an expression more associated with other features such as vocal cord palsy, tracheostomy, tube feeding and GERD which are present in the case reported data, but not well described in HPO. This represents a related but distinct grouping of severe features associated with this condition.

Other smaller clusters identify outlying features that are expressed in individuals or specific families. Many of the outlying features are highly specific facial or anatomical dysmorphisms. While coarse facial features are a noted and core presentation, those with a milder course had clinical reports that articulated these dysmorphisms in much more detail. On a similar note, the reporting of 'talipes' as a condition which seems distinct as a source of variation, appearing as an isolated feature without the typical microcephaly and hypomyelination. However, there was significant heterogeneity in how feet contractures were reported in clinical case reports. In the data pre-processing, a separate feature was created for reported cases with rocker bottom feet or congenital vertical talus versus those with clubfeet/talipes. In practice its not clear how these diagnoses were arrived at, and given the non-specific nature of these features, care needs to be taken in distinguishing between similar manifesting features in the literature where possible. This is particularly clear as rocker bottom feet was found closely related in our analysis to other core features despite the prevalence in HPO being quite low (2/21), suggesting possible under-reporting or confounded reporting of this feature. Another feature that clusters within the core identified features is bilateral cleft lip. This condition has a reported frequency of just 2/21 (Table 8) and may be an emerging feature for a more typical presentation given its close relationship with other classic features. Not all features in the data were included in this analysis, particular if isolated to just one case. In this circumstance, these terms in Table 8 have been labelled as 'Isolated'.

While the number of clusters chosen all represent valid partition of the data, some metrics indicated an optimal selection of 2 clusters. This type of partition is useful when segregating outliers from the rest of the data and is valid in some applications. In this study, a larger number was selected to align the goals of the

15

analysis which was to uncover distinct groups of features.

In terms of patient survival, the first few months show a sharp decline in survival probability confirming early demise in the infant period as a key factor. A small number of cases show much longer survival into childhood and in some cases into the second decade of life. This work identifies a one year survival rate of 0.403 (95% CI: 0.252, 0.644). A naive estimate, using just the number of deceased patients at this time as proportion of the study population is 0.57. This highlights the importance of applying a survival analysis technique that incorporates censoring so as not to erroneously overestimate the survival probability. The current literature suggests the type of gene variant has an impact on survival time, which seems to be supported in this work. Through a comparison of multiple survival curves, stratified by variation type, the limited sample size prevented other meaningful analysis such as log-rank tests or cox proportional hazard modelling which are likely under-powered when comparing multiple groups.

Despite NEDMABA exhibiting a diverse and heterogeneous clinical phenotype, our analysis was able to identify meaningful substructures in this data through the use of dimension reduction techniques (MCA) and cluster analysis. This has isolated the classic NEDMABA phenotype of microcephaly, hypomyelination, arthrogryposis, simplified gyral pattern and developmental delay. Furthermore it has isolated additional distinct groupings of features that are exhibited in patients such as vocal cord paralysis with tracheostomy, swallowing dysfunction, tube feeding and GERD. Furthermore, we were able to quantify the expected baseline survival probabilities for children with this condition as well as quantifying the significant variation in these estimates. Future work is required with the benefit of more cases, to better understand the differences in survival curves for other factors associated with genotype.

5. Appendix

Table 8: A mapping table of phenotype terms from The Human Phenotype Ontology (HPO) with clusters defined in this analysis. Not all clinical features defined in the analysis are present in the HPO terms

HPO Data			Derived Clusters
HPO Term Name	Category	Frequency	**Cluster**[1]
Adducted thumb	Limbs	1/21	Isolated
Simplified gyral pattern	Nervous System	12/18	Core
Hypertonia	Musculature	9/15	Core
Seizure	Nervous System	10/17	Core
Hypotonia	Musculature	4/15	Core
Global developmental delay	Nervous System	7/7	Core
Highly arched eyebrow	Skin, Hair, and Nails	2/21	Isolated
Cerebellar hypoplasia	Nervous System	10/21	Core
Neonatal respiratory distress	Respiratory System	14/17	Adjacent
Bifid uvula	Head and neck	1/21	Isolated
Deep philtrum	Head and neck	1/21	Isolated
CNS hypomyelination	Nervous System	7/19	Core
Lumbar hypertrichosis	Skin, Hair, and Nails	1/21	Isolated
Birth length less than 3rd percentile	Growth	6/11	Core
Hypoplasia of the brainstem	Nervous System	3/20	Isolated
Tented upper lip vermilion	Head and neck	3/21	Isolated
Hypotelorism	Eye	1/21	Isolated
Short palpebral fissure	Head and neck	4/21	Outlying
Bilateral cleft lip	Head and neck	2/21	Core
Diabetes mellitus	Endocrine	2/19	Isolated
Single transverse palmar crease	Skin, Hair, and Nails	1/21	Isolated
Epicanthus	Head and neck	5/21	Outlying
Low anterior hairline	Head and neck	2/21	Outlying
Arthrogryposis multiplex congenita	Connective tissue	17/20	Core
Progressive microcephaly	Nervous System	9/10	Core
Microcephaly	Nervous System	15/21	Core
Thin upper lip vermilion	Head and neck	2/21	Outlying
Thin vermilion border	Head and neck	3/21	Isolated
Intrauterine growth retardation	Growth	12/19	Core
Posteriorly rotated ears	Ear	2/21	Adjacent
Narrow forehead	Head and neck	1/21	Isolated
Sloping forehead	Head and neck	4/21	Isolated
Smooth philtrum	Head and neck	2/21	Isolated
Short philtrum	Head and neck	2/21	Outlying
Premature birth	Prenatal and Birth	7/18	Core
Mandibular prognathia	Head and neck	2/21	Outlying
Wide intermamillary distance	Breast	1/21	Isolated
Depressed nasal bridge	Head and neck	1/21	Outlying
Downslanted palpebral fissures	Head and neck	2/21	Outlying
Short neck	Head and neck	1/21	Isolated
Bulbous nose	Head and neck	2/21	Isolated
Protruding ear	Ear	2/21	Outlying
Rocker bottom foot	Limbs	2/21	Core

Table 9: Assignment of clinical features with their cluster. These clusters represent similar features projected onto the first four MCA principal axes. Clusters that are on opposing sides of an axes contract with each other and account for the largest amount of variation in the data.

cluster	variable
1	polyhydramnios
1	short_palpebral_fissures
1	patent_ductus_arteriosus
1	thin_lips
1	prognathism
1	pectus_excavatum
1	respiratory_infections
2	vocal_cord_palsy
2	respiratory_distress
2	tracheostomy
2	posteriorly_angulated_ears
2	upper_lip_dysmorph
2	tube_feed
2	gerd
2	decreased_craniofacial_ratio
2	short_upturned_nose
2	borderline_small_brainstem
2	enlarged_ventricles
3	epicanthal_folds
3	receding_forehead
3	spastic_tetraparesis
3	intubation
3	low_anterior_hairline
3	short_philtrum
3	hirsutism
4	microcephaly
4	sgp
4	iugr
4	talipes
4	rocker_bottom_feet
4	contractures_hands_deform
4	arthrogryposis_contractures_other
4	cleft_lip
4	coarse_facial_features
4	hypomyelination

cluster	variable
4	cerebellar_hypoplasia
4	small_anterior_fontanel
4	thin_corpus_collosum
4	feeding_swallow_dysfxn
4	hearing
4	seizure
4	agenesis_of_corpus_callosum
4	severe_developmental_delay
4	at_birth_sga
4	lissencephaly_and_cerebellar_hypoplasia
4	eeg_normal
5	hypercapnia
5	inspiratory_stridor
5	oligohydramnios
5	prominent_ears
5	dysmorphic_features
5	tracheo_laryngomalacia

Table 10: Data dictionary of the final data set used for the analysis. This combines the case reports and supplied data for available and references studies. The data has been transformed into a tidy format with formatting applied.

variable	type
study	character
id	character
family	character
individual	factor
deceased	logical
survival_time	numeric
variant_type	factor
locus_1	numeric
locus_2	numeric
gender	factor
ethnicity	character
consanguineous	logical
termination	logical
birth_gestation	numeric
route_vaginal_vs_c_sect	factor
birth_weight	numeric
birth_ofc	numeric
birth_length	numeric
age_at_demise	numeric
age_at_last_follow_up	numeric
microcephaly_ind	factor
sgp_ind	factor
iugr_ind	factor
talipes_ind	factor
rocker_bottom_feet_ind	factor
contractures_hands_deform_ind	factor
vocal_cord_palsy_ind	factor
arthrogryposis_contractures_other_ind	factor

hypercapnia_ind	factor
cleft_lip_ind	factor
inspiratory_stridor_ind	factor
polyhydramnios_ind	factor
coarse_facial_features_ind	factor
oligohydramnios_ind	factor
respiratory_distress_ind	factor
tracheostomy_ind	factor
prominent_ears_ind	factor
posteriorly_angulated_ears_ind	factor
prominent_philtrum_ind	factor
upper_lip_dysmorph_ind	factor
hypomyelination_ind	factor
cerebellar_hypoplasia_ind	factor
short_palpebral_fissures_ind	factor
epicanthal_folds_ind	factor
tube_feed_ind	factor
dysmorphic_features_ind	factor
patent_ductus_arteriosus_ind	factor
small_anterior_fontanel_ind	factor
receding_forehead_ind	factor
tracheo_laryngomalacia_ind	factor
gerd_ind	factor
thin_corpus_collosum_ind	factor
feeding_swallow_dysfxn_ind	factor
seizures_age_started	character
seizure_type_s	character
eeg	character
sleeping_problem	character
behavioral_issues	character
skeletal_changes_scoliosis	numeric
hip_luxation	numeric
development_head_control	character
sitting_unsupported	character
crawling	character
walking	character
speech	character
iq_int_disability	character
cp	character
performance	character
neuro_exam_eye_mvts_nystagmus	character
strabismus	character
feeding_swallow_dysfxn	character
gag	character
tone	character
strength	character
sensory_deficits	character
abnormal_mvts	character
dtr	character
other_neuro_signs	character
sensory_vision	character
hearing	character

other_organs	character
endocrinology	character
functional_tests	character
metabolic_investigation	character
cardiology	character
seizure	numeric
agenesis_of_corpus_callosum	factor
abnormal_pattern_of_cerebral_sulci	factor
left_sided_gallbladder	factor
suspected_brain_stem_dysfunction	factor
spastic_tetraparesis	factor
severe_developmental_delay	factor
acidosis	factor
small_thorax	factor
cyanosis	factor
intubation	factor
apnea	factor
mr_is_reportedly_with_migrational_defect	factor
at_birth_sga	factor
epilepsy_and_apnea	factor
choanal_atresia_surgeries_for_transposition_of_the_great_arteries	factor
interventricular_and_interatrial_defects	factor
smooth_philtrum	factor
thin_lips	factor
bilateral_simian_creases	factor
low_anterior_hairline	factor
short_philtrum	factor
prominent_nose_bridge	factor
depressed_nasal_bridge	factor
prominent_suture_ridges	factor
hypotelorism	factor
hypertrichosis_of_lower_back	factor
decreased_craniofacial_ratio	factor
there_is_overlapping_of_the_sutures	factor
bitemporal_narrowing	factor
short_upturned_nose	factor
short_neck	factor
upward_sweep_of_hair	factor
arched_eye_brows	factor
bulbous_nose	factor
prominent_lower_lip	factor
broad_chin	factor
prognathism	factor
asd	factor
left_multicystic_dysplastic_kidney	factor
anteriorly_placed_anus	factor
dilated_cardiomyopathy	factor
cryptorchidism	factor
hirsutism	factor
intestinal_malrotation	factor
dysphagia	factor
intermittent_sinus_bradycardia	factor

bilateral_staphylomas	factor
lissencephaly	factor
echogenic_choroid_plexus_on_head_us	factor
head_us_mild_ventriculomegaly	factor
brain_parenchyma_appears_homogeneous_without_evidence_of_focal_lesion	factor
hypoplastic_cerebellum	factor
borderline_small_brainstem	factor
abnormal_cerebellar_folia	factor
mild_vermian_hypoplasia	factor
asymmetry_of_lateral_ventricles	factor
lissencephaly_and_cerebellar_hypoplasia	factor
signs_of_hypoxic_encephalopathy	factor
mild_dilation_of_virchow_robins_spaces	factor
pectus_excavatum	factor
widely_spaced_nipples	factor
respiratory_infections	factor
enlarged_ventricles	factor
eeg_normal	numeric
bifid_uvula	factor
under_developed_cerebellar_inferior_vermis	factor

References

, .

Bijarnia-Mahay, S., Somashekar, P.H., Kaur, P., Kulshrestha, S., Ramprasad, V.L., Murugan, S., Sud, S., Shukla, A., 2022. Growth and neurodevelopmental disorder with arthrogryposis, microcephaly and structural brain anomalies caused by bi-allelic partial deletion of smpd4 gene. Journal of human genetics 67, 133–136.

Brock, G., Pihur, V., Datta, S., Datta, S., 2008. clValid: An R package for cluster validation. Journal of Statistical Software 25, 1–22. URL: https://www.jstatsoft.org/v25/i04/.

Chen, X., Sun, X., Hoshida, Y., 2014. Survival analysis tools in genomics research. Human genomics 8, 1–5.

Costa, P.S., Santos, N.C., Cunha, P., Cotter, J., Sousa, N., 2013. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. Journal of aging research 2013.

Datta, S., Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 19, 459–466.

Díaz-Santiago, E., Jabato, F.M., Rojano, E., Seoane, P., Pazos, F., Perkins, J.R., Ranea, J.A., 2020. Phenotype-genotype comorbidity analysis of patients with rare disorders provides insight into their pathological and molecular bases. PLoS genetics 16, e1009054.

Han, L., Benseler, S.M., Tyrrell, P.N., 2018. Cluster and multiple correspondence analyses in rheumatology: paths to uncovering relationships in a sea of data. Rheumatic Disease Clinics 44, 349–360.

Harrington, D.P., Fleming, T.R., 1982. A class of rank test procedures for censored survival data. Biometrika 69, 553–566.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics) 28, 100–108.

Ji, W., Kong, X., Yin, H., Xu, J., Wang, X., 2022. Case report: Novel biallelic null variants of smpd4 confirm its involvement in neurodevelopmental disorder with microcephaly, arthrogryposis, and structural brain anomalies. Frontiers in Genetics 13.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American statistical association 53, 457–481.

Kaufman, L., Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al., 2021. The human phenotype ontology in 2021. Nucleic acids research 49, D1207–D1217.

Kohonen, T., 2012. Self-organizing maps. volume 30. Springer Science & Business Media.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: A package for multivariate analysis. Journal of Statistical Software 25, 1–18. doi:10.18637/jss.v025.i01.

Le Roux, B., Rouanet, H., 2010. Multiple correspondence analysis. volume 163. Sage.

Magini, P., Smits, D.J., Vandervore, L., Schot, R., Columbaro, M., Kasteleijn, E., van der Ent, M., Palombo, F., Lequin, M.H., Dremmen, M., et al., 2019. Loss of smpd4 causes a developmental disorder characterized by microcephaly and congenital arthrogryposis. The American Journal of Human Genetics 105, 689–705.

Monies, D., Abouelhoda, M., Assoum, M., Moghrabi, N., Rafiullah, R., Almontashiri, N., Alowain, M., Alzaidan, H., Alsayed, M., Subhani, S., et al., 2019. Lessons learned from large-scale, first-tier clinical exome sequencing in a highly consanguineous population. The American Journal of Human Genetics 104, 1182–1201.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ravenscroft, G., Clayton, J.S., Faiz, F., Sivadorai, P., Milnes, D., Cincotta, R., Moon, P., Kamien, B., Edwards, M., Delatycki, M., et al., 2021. Neurogenetic fetal akinesia and arthrogryposis: genetics, expanding genotype-phenotypes and functional genomics. Journal of medical genetics 58, 609–618.

Terry M. Therneau, Patricia M. Grambsch, 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York.

Wickham, H., 2014. Tidy data. Journal of Statistical Software 59, 1–23. URL: https://www.jstatsoft.org/index.php/jss/article/view/v059i10, doi:10.18637/jss.v059.i10.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686. doi:10.21105/joss.01686.